Unsupervised Learning

Tianhang Zheng

https://tianzheng4.github.io

Unsupervised Learning vs Supervised Learning

Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response

It is often easier to obtain unlabeled data — from a lab instrument or a computer — than labeled data, which can require human intervention.

The Goals of Unsupervised Learning

We discuss two methods:

Principal components analysis: a tool used for data visualization or data pre-processing before supervised techniques are applied

Clustering: a broad class of methods for discovering unknown subgroups in data.

The Goals of Unsupervised Learning

We discuss two methods:

Principal components analysis: a tool used for data visualization or data pre-processing before supervised techniques are applied

Clustering: a broad class of methods for discovering unknown subgroups in data.

Principal Components Analysis (PCA)

PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.

Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

Principal Components Analysis (PCA)

The first principal component of a set of features is the normalized linear combination of the features

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p$$

that has the largest variance. By normalized, we mean that

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Principal Components Analysis (PCA)

$$\text{var}(x) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

$$\text{cov}(x, y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\begin{bmatrix} \text{Var}(x_1) & \cdots & \text{Cov}(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(x_n, x_1) & \cdots & \text{Var}(x_n) \end{bmatrix}$$

Φ_1 is the eigenvector corresponding to the largest eigenvalue

Do not forget to normalize Φ_1

Clustering Methods

K-means clustering, we seek to partition the observations into a pre-specified number of clusters.

Hierarchical clustering, we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations

K-Means

Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

For instance, if the ith observation is in the kth cluster, then $i \in C_k$.

K-Means

The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.

The problem is formulated as

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \text{WCV}(C_k) \right\}$$

K-Means

Typically use Euclidean distance to measure within-cluster variation

$$\mathrm{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

The objective can be reformulated as

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

K-Means

Typically use Euclidean distance to measure within-cluster variation

$$\mathrm{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

The objective can be reformulated as

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

K-Means

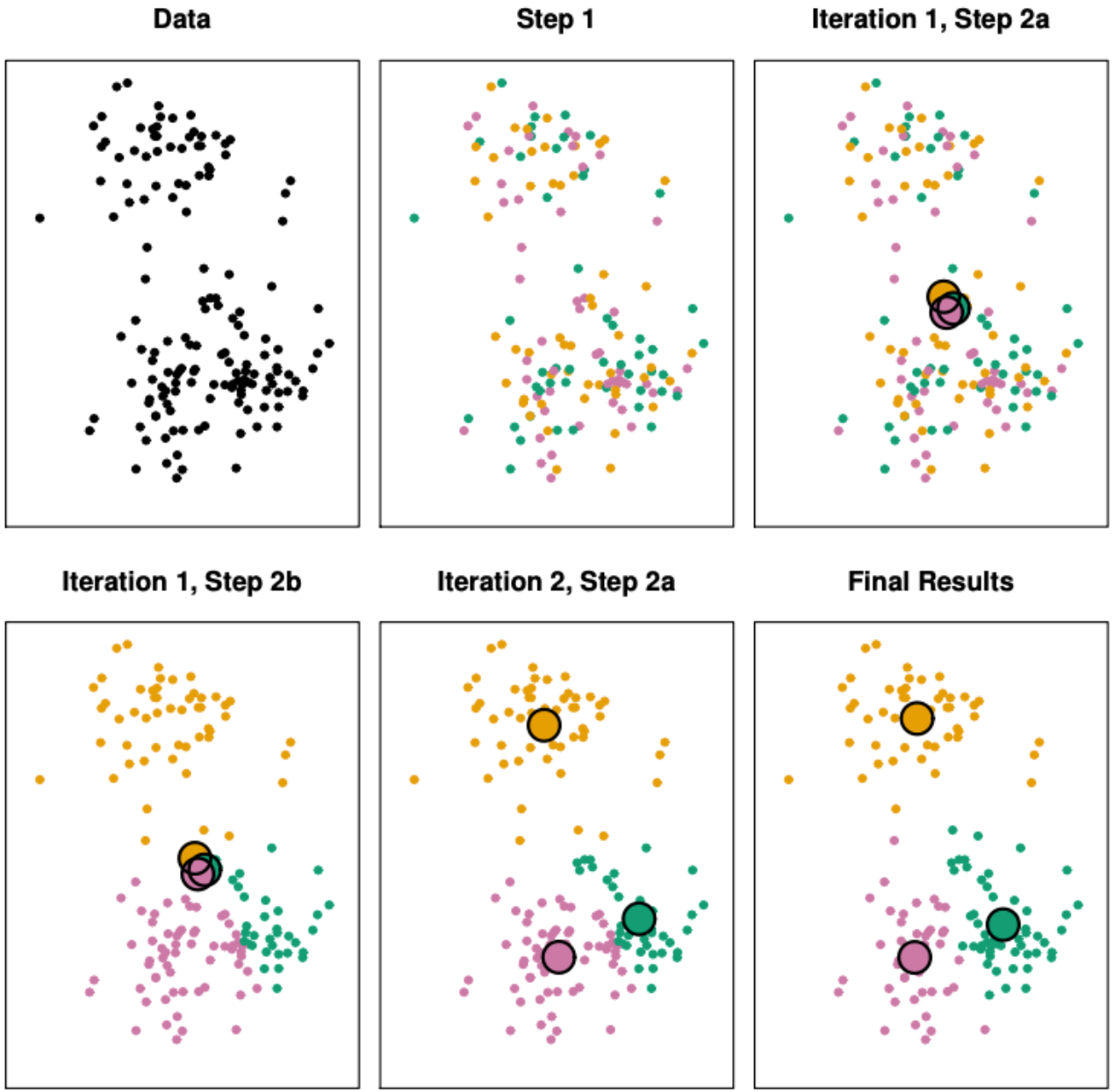
Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

Iterate until the cluster assignments stop changing:

2.1 For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

2.2 Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

K-Means Example



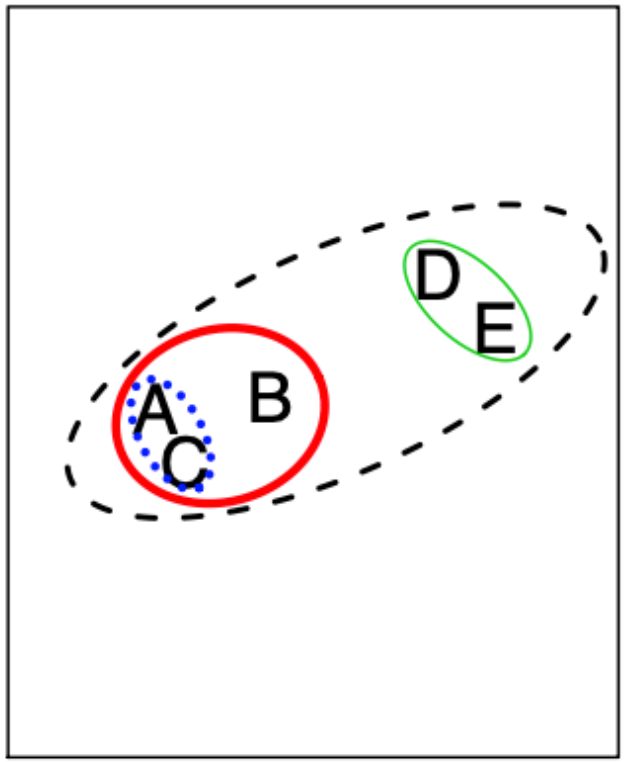
Hierarchical Clustering

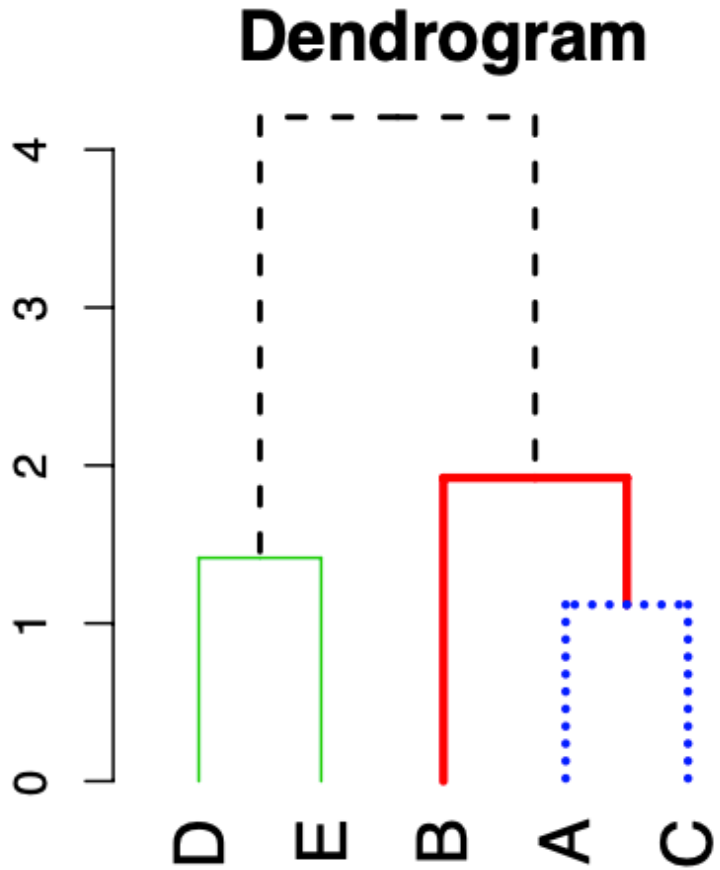
K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage

Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K.

Hierarchical Clustering

Start with each point in its own cluster. Identify the closest two clusters and merge them. Repeat until all points are in a single cluster.





Hierarchical Clustering

How to compute the distance between two clusters?

Minimum distance

Maximum distance

Average distance

Hierarchical Clustering (Number of Clusters)